AVAYA communication

Avaya P333R-LB

Load Balancing Stackable Switch

Load Balancing
Application Guide

May 2001

Communication without boundaries

Table of Contents:

Introduction

Today, businesses rely on the ability to conduct transactions and support customer relationships over the Internet. Delay-free access to information is the key to businesses running smoothly.

Traditional network equipment and design is not suitable to respond to the increasing demand on availability and performance. The two key factors in ensuring you can keep up with your customers' demands are:

The Keys to a Better Network:

- High availability— Users always have access to resources
- Removing bottlenecks— To ensure that servers, firewalls, and other key components are not overloaded and limit network performance.

High Availability—Your clients should not be concerned with your network problems. All they want is 24/7 access without excuses. Load Balancing provides *transparent reliability*—once configured the P333R-LB gives worry-free, availability.

Removing Bottlenecks—High Availability is only half the story; some network element along the way may slow everything down. Any software-based entity in your network, such as a firewall or server can become a bottleneck. By adding more elements in parallel and Load Balancing, you dramatically boost bandwidth.

Load balancing allows you to simply divide demand among multiple resources to ensure that information flows freely.

The applications in this document show how to deploy Load balancing using the P333R-LB to maximize network performance by increasing availability and removing bottlenecks.

What is driving the need for Internet Traffic Engineering?

The Growth of eBusiness Applications: The explosive growth in critical eBusiness applications has created the need to distribute application services over multiple servers for scalability and enhanced availability. The problem with traditional network devices is that many eBusiness applications, such as eCommerce transactions (e.g., shopping carts), search engine sessions, and individualized services (e.g., changing your bank account information online) require that the client communicate with the same server for the entire session. This requires content-aware switching, the ability to examine the packet, and intelligently bind clients to a server using a variety of criteria that may span multiple TCP connections. This goes beyond the capability of traditional L3-4 switches.

Security: Along with increased connectivity and access to the corporate network from extranets and eBusiness applications, comes greater risk to network security. Traditionally, routers and firewalls have been used to protect critical corporate server resources, but as traffic demands grow, their software-based filter processing capabilities hamper access and response. This has created a need for switches that can process an Access Control List (ACL) at hardware-based speeds, while at the same time mitigating external threats from Denial of Service and spoofing attacks.

The P333R-LB

The P333R-LB enhances the P330 stackable switch solution by providing full load balancing and routing functionality in a single switch.

Recent trends in network equipment in general, and switches in particular, have been towards ease of use and deployment.

Avaya leads these trends: the P333R-LB continues the P330 family tradition of being powerful, simple, flexible and offering an unprecedented price/performance ratio.

Flexible

The P333R-LB has the same flexibility as the P330 system: The P333R-LB has both expansion module to allow a variety of uplinks (fiber Ethernet, GE copper fiber or GBIC and even ATM) and a stackability option.

This stackability option allows scalability and investment protection for customers needing more ports and load balancing throughput, increasing investment protection on any network design.

Transparent

Once configured, the P333R-LB "disappears" as far as your users are concerned. As a result, you do not need to perform a large-scale reconfiguration to benefit from increased availability and throughput.

Price/Performance

However good a product is, it must deliver the performance you need at a viable price. The P333R-LB provides all your load balancing needs in a single, compact switch.

With the P333R-LB, you can now afford to implement full Layer 2 and 3 switching and Layer 4 load balancing anywhere in your network. Now applications such as Server load balancing, Firewall Load Balancing, Application Redirection and Policy-based load balancing have become cost-effective.

The P333R-LB provides all this in one switch. Compare this with other solutions that require several expensive switches to perform the same functionality.

Application 1—Server Load Balancing

Server load balancing allows you to provide high availability and overcome bottlenecks caused by limitations in the servers.

Configuring mirrored servers via load balancing means that, if one server is taken off-line for any reason, users will not experience any delays or interruptions. You can extend this availability to a number of levels, from servers to server farms to load balancing switches.

Load balancing can also overcome bottlenecks created by "server session overload." Without load balancing, the number of sessions that the server can handle may become insufficient for all the users who are interacting with your organization.

By intelligently dividing the workload among your servers, you ensure that users can continue to interact in a bottleneck-free environment.

Installing one or more P333R-LB stackable load balancing switches as the server farm switch allows you to easily implement Server load balancing. The server farm appears as a single Virtual IP (VIP) address to the outside world, so you can implement multiple servers for increased throughput and redundancy with no disruption.

Load Balancing Schemes: You can choose the load balancing scheme according to your needs. The P333R-LB offers Round Robin, Hash and Response Time-based schemes.

- Round Robin—This scheme directs traffic to the next server in the list.
- Hash—This statistically-based scheme is based on a pre-defined algorithm.
- Response Time-based—This scheme directs traffic to the least busy server.

Persistency. Persistency is important since it ensures that users receive consistent and uninterrupted responses to their requests.

Persistency is especially important in eCommerce where persistency of a session with the same server ensures that transactions between the user and a specific server will not be lost. Persistency ensures that the SSL encryption session will not be cut by moving to a different server. The P333R-LB offers two persistency schemes: Hash, Source / destination IP-based (and destination Port), and activity timer-based.

Health Checking: In order to make the load balancing more effective in terms of availability, we have incorporated a health checking mechanism into the load balancer that ensures that traffic is only directed to "healthy" servers. This vital component of load balancing is performed by "pinging" the servers and checking for a response.

Graceful Shutdown: This allows you to take down a server to perform maintenance without your users feeling a thing. Old sessions continue until they are completed, and new sessions are directed to another server. Once all old sessions are completed, you can take down the server.

The Load Balancing Process

The P333R-LB selects the best available physical server upon receiving a new session request based on the following:

- Server health (based on application health, not just server status)
- Number of open sessions
- Round Robin
- Hash
- Persistence: Subsequent packets in the same session sent to the same server until the session terminates.
- Session address translation performed on every packet

This process is transparent both to the user and to the server, so no configuration is required (unlike proxy servers, for example).

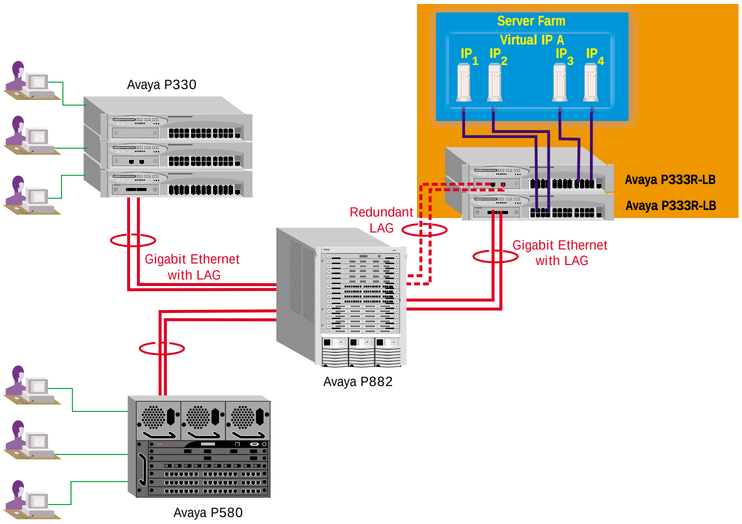
Unlike simple load-balancers, the Avaya load balancing solution supports multiple persistence methods. This enables the P333R-LB to fully support eBusiness applications that require continuous communication between the client and server throughout a transaction session. The following persistence methods are supported:

- IP address hashing
- Source address with inactivity timer

High Availability with Cajun

The special configuration below shows how each element is fully backed up to provide maximum uptime:

- Server fails—The P333R-LB senses the loss of the server and directs all new sessions to the other servers that are available in the farm.
- P333R-LB fails—enhanced VRRP ensures transparent switchover to the second switch. A special topology ensures the server connection to the network.
- Uplink fails—STA and port redundancy bring the backup LAG on-line instantaneously.



*future

Application 2—Firewall Load Balancing

Firewalls can inherently constitute a bottleneck since they are software-based. There is also the issue of transparent availability: firewalls can act as a single point of failure, causing severe problems with Internet access.

The P333R-LB can be used to load-balance across multiple routers and firewalls and overcome these problems.

Implementing the P333R-LB removes the bottleneck since the load is distributed at hardware speed over multiple firewalls.

To ensure availability, the P333R-LB switches on both sides of the firewall perform continuous health checks on the links to the firewall, the firewall itself, and each other.

The two applications below show firewall load balancing in systems both with and without NAT (Network Address Translation).

No single point of failure. The doubling up of the key components ensures that communication between the LAN and Internet is always maintained.

If one P333R-LB fails, due to VRRP, the second switch can instantaneously take over all load balancing functions. If a firewall fails, then the P333R-LB will transparently redirect all traffic through the second firewall.

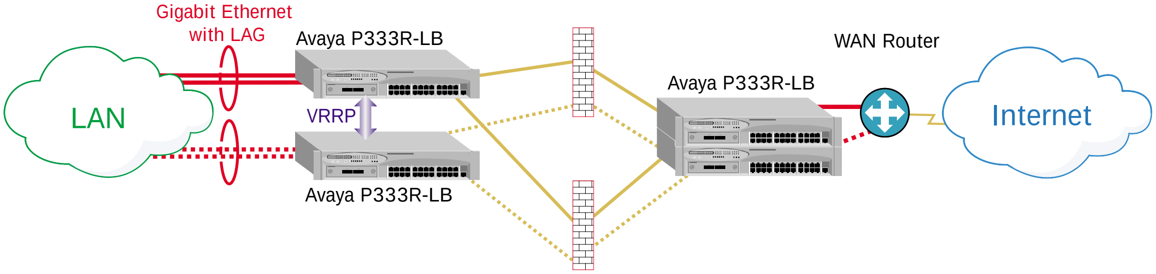
"No NAT" Application. In this case, there are pairs of P333R-LB switches on each side of the firewalls. This is necessary since sessions must travel across the same firewall. If the session is sent to the second firewall, it will be disconnected by the "statefull" firewall. It is therefore important to have the same load balancing decisions on both sides of the firewall.

NAT Application. In this case, P333R-LB switches are only required on the LAN side of the firewalls. The session traffic coming from the Internet will have the specific IP address of the firewall from which the session started.

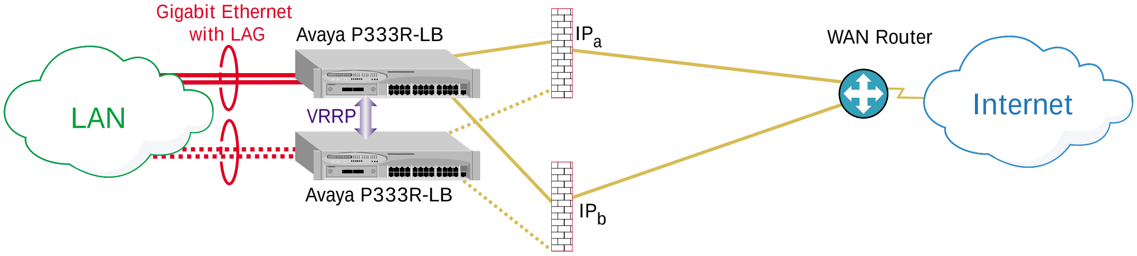
- NAT (Network Address Translation)—allows you to use any IP address within your organization while only using legal IP addresses outside.

  This is useful when you have a limited number of legal IP addresses (e.g., for ISPs).

Firewall Load Balancing – no NAT



Firewall Load Balancing – with NAT



Application 3 — Application Redirection

Application redirection is the diversion of traffic from certain applications to a different direction than the original one without the user / source being aware or requiring any reconfiguring.

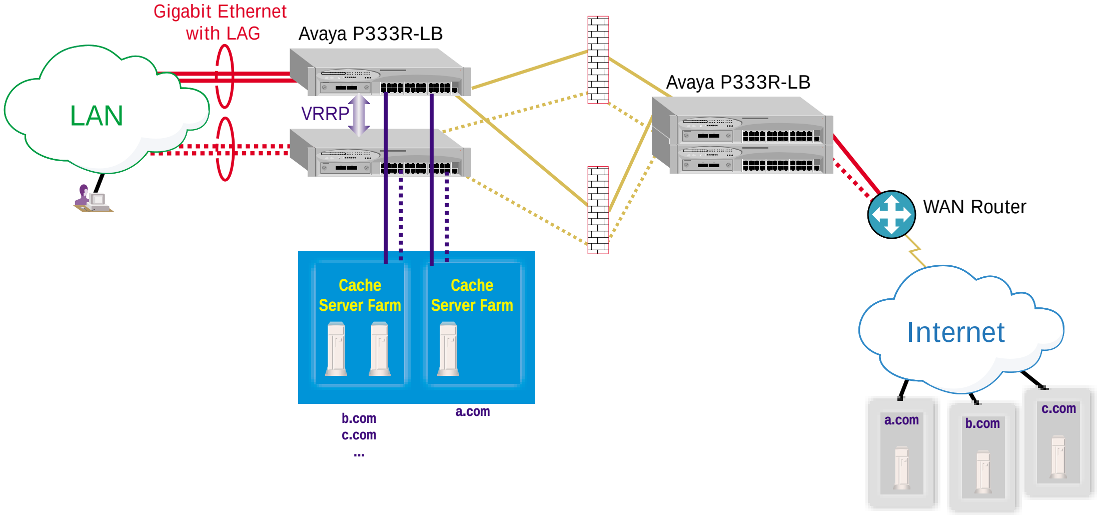
Cache Redirection: One of the most common uses of application redirection is "cache redirection." Today, many organizations use proxy servers for Internet access. This approach, however, has problems which makes it less than ideal:

- It is only possible to deploy a single proxy server which is both a bottleneck and a single point of failure.
- The network manager needs to configure each user to access the Internet via the proxy server IP.
- Users can bypass the proxy without supervision.

Solution to the Problem: With cache redirection, the user does not need to configure anything: they are convinced they are accessing the Internet directly. Load balancing identifies Internet traffic, "snatches" the packets and diverts them to the cache.

The cache server checks whether the information is in its memory. If it is not, the cache server retrieves the information from the Internet. The cache server then forwards the information to the user who is unaware of the entire process.

It is possible to deploy more than one server and to implement Server load balancing among them *(see also Application 1 — Server Load Balancing).*



Application 4 — Policy-Based Load Balancing

Policy-based load balancing allows you to provide tiered services. Preferred users receive better service / access and you can charge those customers according to the level of service they receive. This process of maximizing the use of current resources is transparent to the user.

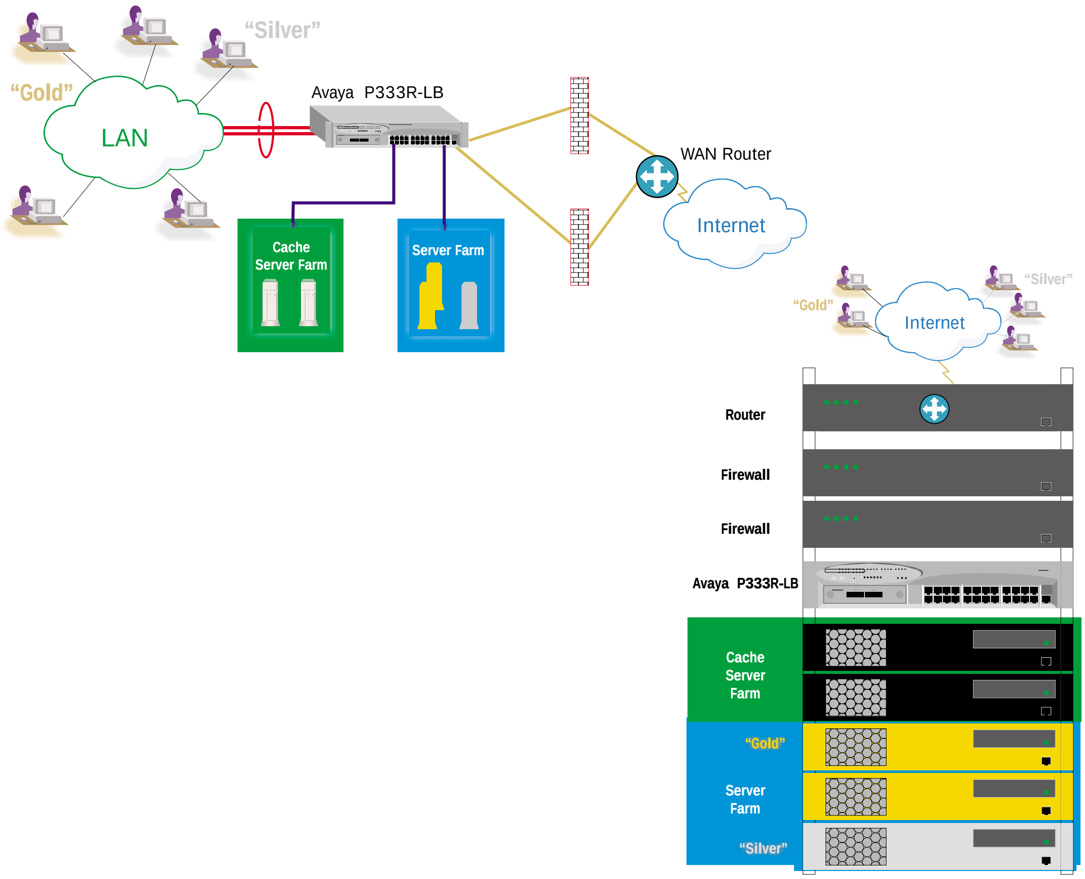
Implementation: The first stage is to define groups of users and servers (gold, silver, etc.). The servers provide identical services. "Gold" users are directed to the "Gold" servers, "Silver" users are directed to the "Silver" servers, etc.

- Users receive service according to their level: "gold," "silver" or "bronze."

Advantages: Other load balancers on the market redirect traffic based only on a division of users and servers to "Gold" and "Silver" groups. However, Avaya's policy-based load balancing gives the best use of current resources by continuously checking the status of the "Gold" and "Silver" groups. If the "Gold" group is more loaded than the "Silver" group then the "Gold" user will receive service from the "Silver" group. Thus "Gold" users will always receive the best service.

Policy-Based Load Balancing and Service Providers

The same application can be implemented with cache servers, so Service Providers can give the best service to preferred customers. The customers only notice improved service—without any reconfiguration and complications.



AVAYA
communication

For a contact in your area, go to: www.avaya.com/contactus